

Model-based clustering for conditionally correlated categorical data

Matthieu Marbac*
Inria & DGA

Christophe Biernacki†
University Lille 1 & CNRS & Inria

Vincent Vandewalle‡
EA 2694 University Lille 2 & Inria

July 11, 2014

Abstract

An extension of the latent class model is presented for clustering categorical data by relaxing the classical “class conditional independence assumption” of variables. This model consists in grouping the variables into inter-independent and intra-dependent blocks, in order to consider the main intra-class correlations. The dependency between variables grouped inside the same block of a class is taken into account by mixing two extreme distributions, which are respectively the independence and the maximum dependency. When the variables are dependent given the class, this approach is expected to reduce the biases of the latent class model. Indeed, it produces a meaningful dependency model with only a few additional parameters. The parameters are estimated, by maximum likelihood, by means of an EM algorithm. Moreover, a Gibbs sampler is used for model selection in order to overcome the computational intractability of the combinatorial problems involved by the block structure search. Two applications on medical and biological data sets show the relevance of this new model. The results strengthen the view that this model is meaningful and that it reduces the biases induced by the conditional independence assumption of the latent class model.

Keywords: Categorical data; Clustering; Correlation; Expectation-Maximization algorithm; Gibbs sampler; Mixture model; Model selection.

*Matthieu Marbac is a Ph.D student at Inria Lille and DGA (Email: matthieu.marbac-lourdelle@inria.fr)

†Christophe Biernacki is a Professor at University Lille 1, CNRS and Inria Lille (Email: Christophe.Biernacki@math.univ-lille1.fr)

‡Vincent Vandewalle is an Associate Professor at University Lille 2, EA 2694 and Inria Lille (Email: vincent.vandewalle@univ-lille2.fr)

1 Introduction

Nowadays practitioners often face very large data sets, which are difficult to analyze directly. In this context, clustering (Jajuga *et al.* , 2002) is an important tool providing a partition among individuals. Other approaches cluster both individuals and variables simultaneously (Govaert & Nadif, 2003). Furthermore, with the increasing number of variables at hand, the risk of observing correlated descriptors, even within the same class, is often high. In view of these difficulties, the practitioner can choose between two approaches. The first approach consists in performing a selection among the observed variables (Maugis *et al.* , 2009) in order to extract uncorrelated data, thereby losing some potentially crucial information. The second approach consists in applying a method for modeling the conditional dependencies on the whole set of variables.

Clustering methods can be divided into two kinds of approaches: geometric ones based on the distances between individuals and probabilistic ones modelling the data generation process. Geometric methods are generally faster than probabilistic ones, but they are often quite sensitive to the choice of distance between individuals. Furthermore, as the probabilistic tools are not available for these methods, difficult questions, e.g. the selection of the number of clusters cannot be addressed rigorously. For categorical data, geometric methods either define a metric in the initial variables space e.g. *k-means* (Huang *et al.* , 2005), or compute their metric on the axes of the multiple correspondence analysis (Chavent *et al.* , 2010; Guinot *et al.* , 2001).

Many geometric approaches can also be interpreted in a probabilistic way. Thus, for the continuous data, the classical *k-means* algorithm can be identified as an homoscedastic Gaussian mixture model (Banfield & Raftery, 1993; Celeux & Govaert, 1995) with equal proportions. For the categorical variables, Celeux & Govaert (1991) show that the CEM algorithm (McLachlan & Krishnan, 1997), applied to a classical latent class model, maximizes a classical information criteria close to a χ^2 metric. Other links between the two approaches are described in Govaert (2010), Chapter 9. Let us now introduce our proposal for this problem.

In the categorical case, the *latent class model* also known as naive Bayes belongs to the folklore (Goodman, 1974; Celeux & Govaert, 1991). In this article, we refer to this model as the conditional independence model (further denoted by CIM). Classes are explicitly described by the probability of each modality for each variable under the conditional independence assumption.

The sparsity of the model, implied by this assumption, is a great advantage since it restricts the curse of dimensionality. CIM has obtained quite good results in practice in different areas (Hand & Yu, 2001), e.g. behavioral science (Reboussin *et al.*, 2006) and medicine (Strauss *et al.*, 2006). However, CIM may suffer from severe biases when the data are intra-class correlated. For instance, an application presented by Van Hattum & Hoijsink (2009) shows that CIM over-estimates the number of clusters when the conditional independence assumption is violated. For a long time, people have tried to relax the conditional independence assumption by modeling conditional interactions between variables using an additive model (Harper, 1972). The main drawback of this approach is that its number of parameters becomes huge, making their estimation intractable.

Some other methods take into account the intra-class dependencies as *mixtures of Bayesian networks* (Cheng & Greiner, 1999). Conditionally on each class, a directed acyclic graph is built with a set of nodes representing each variable. However, if no constraint is added, the network estimation is also quite complex. By constraining the network to be a tree, the model selection and the parameter estimation can be easily performed. Moreover the correlation model enjoys great flexibility. The extension of the dependency tree of Chow & Liu (1968) was done by Friedman *et al.* (1997) for the supervised classification and by Meila & Jordan (2001) for the clustering. However the main problem of these models is that they too often require too many parameters.

When covariates are available, the conditional dependencies between the categorical ones can be modeled by a logistic function (Formann, 1992; Reboussin *et al.*, 2008). By assuming that these covariates are unobserved, the *multilevel latent class model* (Vermunt, 2003, 2007) naturally incorporates the intra-class dependencies. This model has connections with the approach of Qu *et al.* (1996) where the intra-class dependencies are modeled by a latent continuous variable with a probit function. The *hybrid model* (Muthén, 2008) is more general approach in which, for each class, a factor analysis model is fitted to either all categorical variables or to those categorical variables having dependencies. Recently, Gollini & Murphy (2013) proposed the *mixture model of latent traits analyzers* which assumes that the distribution of the categorical variables depends on both a categorical latent variable (the class) and many continuous latent traits variables. The parameter estimation is also a difficult point which is solved via a variational approach. All these models consider the intra-class dependencies, but their main drawback is that these dependencies must be interpreted among relations with a latent variable. As a result, pertinent interpretation

can be difficult.

The log-linear models (Agresti, 2002; Bock, 1986) were originally proposed to model the log-probability of each individual by selecting interactions between variables. The most general mixture model is the *log-linear mixture model* as it is able to incorporate many forms of interactions. It has been used at least since Hagenaars (1988). Espeland & Handelman (1989) used it to cluster radiographic cross-diagnostics and Van Hattum & Hoijsink (2009) in a market segmentation problem. However this model family is huge and the model selection is a real challenge. In the literature, authors often require the modeled interactions ahead of time. Another option is to perform a deterministic search e.g. the *forward* method which is sub-optimal. Furthermore, the number of parameters to estimate increases with the conditional modalities interactions, thus implying potential over-fitting and more difficult interpretation. The latent class model (LCM) can be seen as a particular log-linear mixture model, in which interactions are discarded. Our aim is to present a version of the log-linear mixture model which takes into account the interactions of order one or more while keeping the number of unknown parameters to a reasonable amount.

We propose to extend the classical latent class model (LCM) for categorical data, by a new latent class model which relaxes the conditional independence assumption. We refer to this new model as the *conditionally correlated model* (denoted by CCM). This model is a parsimonious version of the log-linear mixture model, and thus benefits from its interpretative power. Furthermore, we propose a Bayesian approach to automatically perform model selection.

The CCM model groups the variables into conditionally independent blocks given the class. The main intra-class dependencies are thus shown by the repartition of the variables into blocks. This approach, allowing modeling of the main conditional interactions, was first proposed by Jorgensen & Hunt (1996) in order to cluster continuous and categorical data. For CCM, each block follows a particular dependency distribution which corresponds to our main contribution. This distribution consists in a bi-component mixture of an *independence* and a *maximal dependency* distribution according to the Cramer's V criterion. This specific distribution of the blocks allows summarizing the conditional dependencies of the variables with only one continuous parameter: the maximum dependency distribution proportion. Thus, the model underlines the main conditional dependencies and their strength.

The new model is a two-degree parsimonious version of a log-linear mixture model. A first

degree of parsimony is introduced by grouping the variables which are conditionally dependent into the same block. This repartition of the variables per blocks defines the interactions considered by the model for each class. Moreover, the strength of the correlation is reflected by the proportion of maximum dependency distribution. A second degree of parsimony is induced by the specific distribution of the blocks. As for all log-linear mixture models, the selection of the pertinent interactions is a combinatorial problem. We propose to perform this model selection via a Gibbs sampler in order to overcome the enumeration of all the models. Thus, this general approach could also select the interactions of any log-linear mixture model.

This paper is organized as follows. Section 2 reviews the principles of the latent class model. Section 3 presents the new mixture model taking into account the intra-class correlations. Section 4 is devoted to parameter estimation in the case where the number of classes and the blocks of variables are supposed to be known. Section 5 presents a Gibbs algorithm for avoiding the combinatorial difficulties inherent to block selection. Section 6 presents results on simulated data. Section 7 presents a comparison between two main model-based clustering approaches and our proposal on a classical medical data set, and presents another application on a larger real data set. A tutorial of the R package `Clustericat`¹ which performs the model selection and the estimation of the parameters of CMM is given with the first application (see Appendix A). Section 8 presents our conclusions.

2 Classical models

2.1 Latent class model: intra-class independence of variables

Observations to be classified are described with d discrete variables $\mathbf{x} = (\mathbf{x}^1, \dots, \mathbf{x}^d)$ defined on the probabilistic space \mathcal{X} . Each variable j has m_j response levels with $m_j \geq 2$ and is written $\mathbf{x}^j = (x^{j1}, \dots, x^{jm_j})$ where $x^{jh} = 1$ if variable j takes modality h and $x^{jh} = 0$ otherwise. In the standard latent class model (CIM), the variables are assumed to be *conditionally independent* knowing the latent cluster. Furthermore, data are supposed to be drawn independently from a mixture of g multivariate multinomial distributions with probability distribution function (pdf)

¹The R package `Clustericat` is available on Rforge website at the following url: https://r-forge.r-project.org/R/?group_id=1803

$$p(\mathbf{x}; \boldsymbol{\theta}) = \sum_{k=1}^g \pi_k \dot{p}(\mathbf{x}; \boldsymbol{\alpha}_k) \quad \text{with} \quad \dot{p}(\mathbf{x}; \boldsymbol{\alpha}_k) = \prod_{j=1}^d \prod_{h=1}^{m_j} (\alpha_k^{jh})^{x^{jh}}, \quad (1)$$

with $\boldsymbol{\theta} = (\pi_1, \dots, \pi_g, \boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_g)$, π_k being the proportion of the component k in the mixture where $\pi_k > 0$ and $\sum_{k=1}^g \pi_k = 1$, and $\boldsymbol{\alpha}_k = (\alpha_k^{jh}; j = 1, \dots, d; h = 1, \dots, m_j)$ where α_k^{jh} denotes the probability that variable j has level h if the object is in cluster k and satisfies the two following constraints: $\alpha_k^{jh} > 0$ and $\sum_{h=1}^{m_j} \alpha_k^{jh} = 1$.

The classical latent class model is much more parsimonious than the saturated log-linear model, which requires $(\prod_j m_j) - 1$ parameters, since it only requires ν_{CIM} parameters with

$$\nu_{\text{CIM}} = (g - 1) + g \sum_{k=1}^g (m_j - 1). \quad (2)$$

Its maximum likelihood estimator is easily computed via an EM algorithm (McLachlan & Krishnan, 1997). For the cluster analysis, the mixture identifiability up to a permutation of the class is generally necessary (McLachlan & Peel, 2000). However, there are mixtures, such as the products of Bernoulli distributions, which are not identifiable but which produce good results in several applications. In order to relax this too stringent concept of identifiability, the notion of generic identifiability was introduced by Allman *et al.* (2009): a model is generically identifiable if it is identifiable except for a subset of the parameter space with Lebesgue measure zero.

2.2 Latent class model extension: intra-class independence of blocks

Despite its simplicity, the latent class model s to good results in many situations (Hand & Yu, 2001). However, in the case of intra-correlated variables, it can entail severe biases in the partition estimation and it may also overestimate the number of clusters. In order to reduce these biases, a classical extension of the latent class model was introduced by Jorgensen & Hunt (1996) for conditionally correlated mixed data. This model is implemented in the Multimix software (Hunt & Jorgensen, 1999).

It considers that *conditionally* on the class k , variables are grouped into B_k *independent blocks* and that each block follows a specific distribution. The repartition in blocks of the variables determines a partition $\boldsymbol{\sigma}_k = (\boldsymbol{\sigma}_{k1}, \dots, \boldsymbol{\sigma}_{kB_k})$ of $\{1, \dots, d\}$ in B_k disjoint non-empty subsets where $\boldsymbol{\sigma}_{kb}$ represents the subset b of variables in the partition $\boldsymbol{\sigma}_k$. This partition defines $\mathbf{x}^{\{kb\}} = \mathbf{x}^{\boldsymbol{\sigma}_{kb}} =$

$(\mathbf{x}^{\{kb\}j}; j = 1, \dots, d^{\{kb\}})$ which is the subset of \mathbf{x} associated to σ_{kb} . The integer $d^{\{kb\}} = \text{card}(\sigma_{kb})$ is the number of variables in block b of component k and $\mathbf{x}^{\{kb\}j} = (x^{\{kb\}jh}; h = 1, \dots, m_j^{\{kb\}})$ corresponds to variable j of block b for component k with $x^{\{kb\}jh} = 1$ if the individual takes modality h for variable $\mathbf{x}^{\{kb\}j}$ and $x^{\{kb\}jh} = 0$ otherwise and where $m_j^{\{kb\}}$ represents the number of modalities of $\mathbf{x}^{\{kb\}j}$. Note that different repartitions of the variables into blocks are allowed for each component and they are grouped into $\sigma = (\sigma_1, \dots, \sigma_g)$.

For each component k , each block b follows a specific parametric distribution denoted as $p(\mathbf{x}^{\{kb\}}; \theta_{kb})$ where θ_{kb} groups the parameters of this distribution. The model pdf is written as

$$p(\mathbf{x}; \sigma, \theta) = \sum_{k=1}^g \pi_k p(\mathbf{x}; \sigma_k, \theta_k) \quad \text{with} \quad p(\mathbf{x}; \sigma_k, \theta_k) = \prod_{b=1}^{B_k} p(\mathbf{x}^{\{kb\}}; \theta_{kb}), \quad (3)$$

where θ is redefined as $\theta = (\pi_1, \dots, \pi_g, \theta_1, \dots, \theta_g)$ with $\theta_k = (\theta_{k1}, \dots, \theta_{kB_k})$. Figure 1 presents an example of the distribution with conditional independent blocks for a mixture with two components described by five variables. Blank cells indicate that the intra-class correlation is neglected and black cells indicate that this correlation is taken into account. For instance, Figure 1.a illustrates the distribution of the class 1 where two blocks ($B_1 = 2$) are considered. More precisely, the first block is composed by the first two variables ($\sigma_{11} = \{1, 2\}$) and the second block is composed by the last three variables ($\sigma_{12} = \{3, 4, 5\}$). Note that the classical latent class model with conditional independence, would be represented by white cells off the diagonal and black cells on the diagonal.

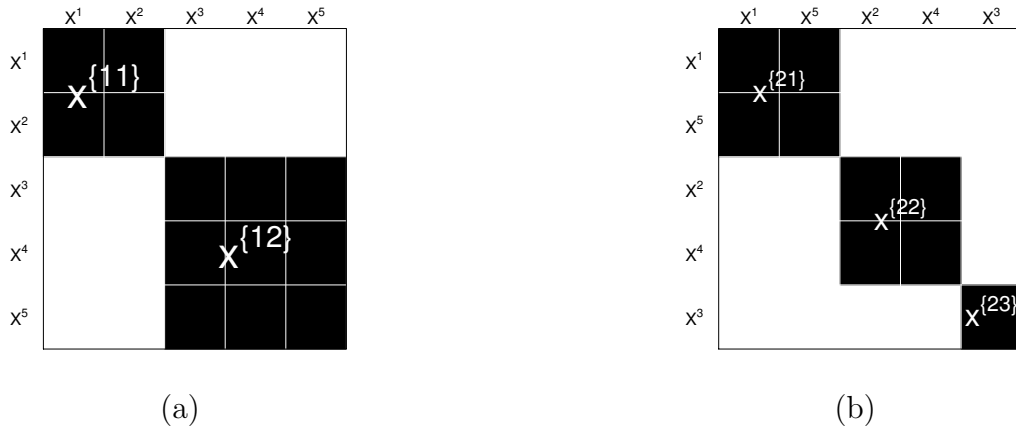


Figure 1: Example of the conditional independent blocks mixture model with $g = 2$ and $d = 5$, (a) $k = 1$, $B_1 = 2$ and $\sigma_1 = (\{1, 2\}, \{3, 4, 5\})$, (b) $k = 2$, $B_2 = 3$ and $\sigma_2 = (\{1, 5\}, \{2, 4\}, \{3\})$.

This approach is very general, since any distribution can be chosen for each block as soon as it

is different from the distribution of independence. The mixture model by conditional independent blocks is a parsimonious version of the log-linear mixture model. Indeed, the distribution of variables in blocks determines which interactions are modeled. The interactions between variables of different blocks will be zero and those between variables of the same block can be modeled by the specific distribution of the block. The limiting case of this model where $B_k = d$ for each class is equivalent to the latent class model with the conditional independence assumption.

The generic identifiability of the mixture model with conditionally independent blocks follows, under specific constraints, from Theorem 4 of Allman *et al.* (2009) by assuming that the distribution of each block is itself identifiable. This proof is given in Appendix A of Marbac *et al.* (2013).

3 Intra-block parsimonious distribution

The goal is now to define a parsimonious distribution for each block that takes into account the dependency between variables. Furthermore, the parameters of the distribution inside the block must be meaningful for the practitioner. In this context, we propose to model the distribution of each block by a mixture of the extreme distributions according to the Cramers V criterion computed on all the variable couples. The model results in a bi-component mixture between an independence distribution and a maximum dependency distribution which can be easily interpreted by the user. The maximum dependency distribution is introduced first. The resulting conditional correlated model (CCM) is also defined as a block model extension of the latent class model where the distribution inside the block is modeled by this bi-component mixture. Remark: Without loss of generality, the variables are considered as ordered by decreasing number of modalities in each block: $(k, b) \text{ mkb } j \text{ mkb } j$

The goal is now to define a parsimonious distribution for each block that takes into account the dependency between variables. Furthermore, the parameters of the distribution inside the block must be meaningful for the practitioner. In this context, we propose to model the distribution of each block by a mixture of the extreme distributions according to the Cramers V criterion computed on all the variable couples. The model results in a bi-component mixture between an independence distribution and a maximum dependency distribution which can be easily interpreted by the user. The maximum dependency distribution is introduced first. The resulting conditional

correlated model (CCM) is also defined as a block model extension of the latent class model where the distribution inside the block is modeled by this bi-component mixture.

Remark: Without loss of generality, the variables are considered as ordered by decreasing number of modalities in each block: $\forall(k, b) m_j^{\{kb\}} \geq m_{j+1}^{\{kb\}}$ where $j = 1, \dots, d^{\{kb\}} - 1$.

3.1 Maximum dependency distribution

The maximum dependency distribution is defined as the “opposite” distribution of independence according to the Cramer’s V criterion computed on all the variable couples. Indeed, this latter minimizes this criterion while the maximum dependency distribution maximizes it. Under this distribution, the modality knowledge of one variable provides the maximum information on all the subsequent variables. Note that it is a non-reciprocal functional link between variables. Indeed, if $\mathbf{x}^{\{kb\}}$ arises from this distribution, the knowledge of the variable having the largest number of modalities determines exactly the others but the reverse does not necessarily apply. So, this distribution defines successive surjections from the space of $x^{\{kb\}j}$ to the space of $x^{\{kb\}j+1}$ with $j = 1, \dots, d^{\{kb\}} - 1$ (recall that the variables are ordered by decreasing number of modalities in each block). In fact, it is a reciprocal functional link only when $m_j^{\{kb\}} = m_{j+1}^{\{kb\}}$.

Since the first variable determines the other ones, this distribution is defined by a product between the multinomial distribution of the first variable parametrized by $\boldsymbol{\tau}_{kb} = (\tau_{kb}^h; h = 1, \dots, m_1^{\{kb\}})$ with $\tau_{kb}^h \geq 0$ and $\sum_{h=1}^{m_1^{\{kb\}}} \tau_{kb}^h = 1$, and the product between the conditional distributions defined as specific multinomial distributions. So, conditionally on $x^{\{kb\}1h} = 1$, $\mathbf{x}^{\{kb\}j}$ is deterministic for each $j = 2, \dots, d^{\{kb\}}$. Indeed, in such a case, $\mathbf{x}^{\{kb\}j}$ follows a specific multinomial distribution whose parameters are 0 and 1. More precisely, this distribution is parametrized by $\boldsymbol{\delta}_{kb}^{hj} = (\delta_{kb}^{hjh'}; h' = 1, \dots, m_j^{\{kb\}})$ with the following constraints defining the successive surjections: $\delta_{kb}^{hjh'} \in \{0, 1\}$, $\sum_{h'=1}^{m_j^{\{kb\}}} \delta_{kb}^{hjh'} = 1$ (multinomial distribution) and $\sum_{h=1}^{m_1^{\{kb\}}} \delta_{kb}^{hjh'} \geq 1$ (surjections).

By denoting $\boldsymbol{\delta}_{kb} = (\boldsymbol{\delta}_{kb}^{hj}; h = 1, \dots, m_1^{\{kb\}}; j = 2, \dots, d^{\{kb\}})$, the pdf of maximum dependency distribution is defined as:

$$\begin{aligned} p(\mathbf{x}^{\{kb\}}; \boldsymbol{\tau}_{kb}, \boldsymbol{\delta}_{kb}) &= p(\mathbf{x}^{\{kb\}1}; \boldsymbol{\tau}_{kb}) \prod_{j=2}^{d^{\{kb\}}} p(\mathbf{x}^{\{kb\}j} | \mathbf{x}^{\{kb\}1}; \{\boldsymbol{\delta}_{kb}^{hj}\}_{h=1, \dots, m_1^{\{kb\}}}) \\ &= \prod_{h=1}^{m_1^{\{kb\}}} \left(\tau_{kb}^h \prod_{j=2}^{d^{\{kb\}}} \prod_{h'=1}^{m_j^{\{kb\}}} (\delta_{kb}^{hjh'})^{x^{\{kb\}jh'}} \right)^{x^{\{kb\}1h}}. \end{aligned} \quad (4)$$

Figure 2 shows two examples of the maximum dependency distributions. The probabilities of the joint distribution are represented by the area of dark boxes. Notice that δ_{kb} defines the position where the probabilities are non-zero (location of a dark boxes) and τ_{kb} defines the probabilities of this non-zero cells (area of the dark boxes). For the example illustrated in Figure 2.a, the probability that the first variable takes the modality one is 0.1 ($\tau_{11}^1 = 0.1$). Moreover, conditionally on $x^{\{11\}11} = 1$, variable $x^{\{11\}2}$ is deterministic since the probability that $x^{\{11\}21} = 1$ given $x^{\{11\}11} = 1$ is one.

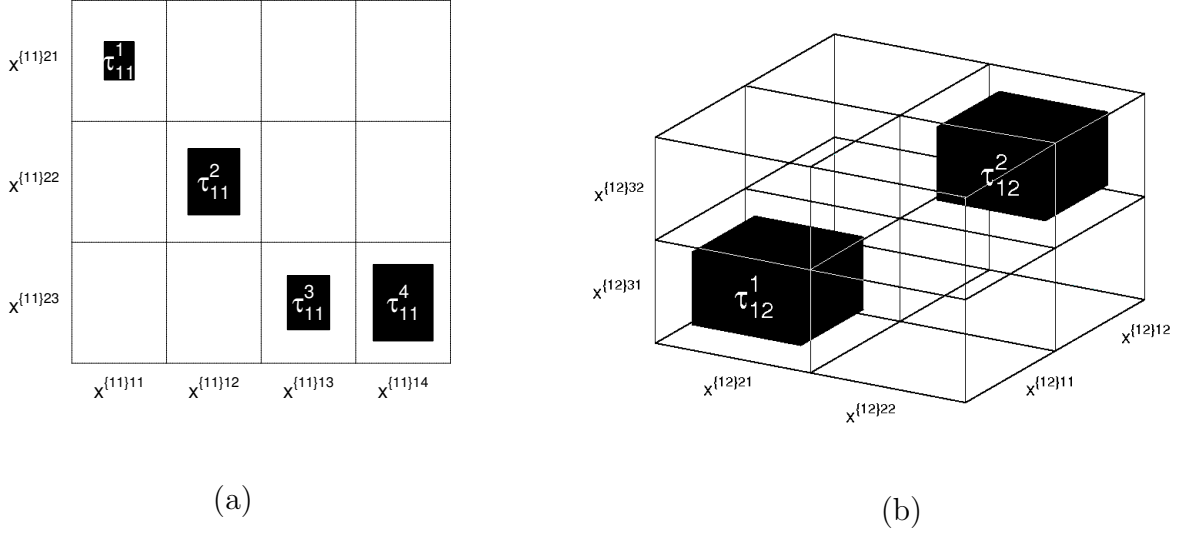


Figure 2: Two examples of the maximum dependency distributions for the first component of the mixture illustrated by Figure 1(a). (a) The first block is displayed with $m^{\{11\}1} = 4$, $m^{\{11\}2} = 3$, $\delta_{11}^{h1h} = 1$ for $h = 1, 2, 3$, $\delta_{11}^{413} = 1$ and $\tau_{11} = (0.1, 0.3, 0.2, 0.4)$; (b) The second block is displayed with $m^{\{12\}1} = m^{\{12\}2} = m^{\{12\}3} = 2$, $\delta_{12}^{hjh'} = 1$ iff $(h = h')$ and $\tau_{12} = (0.5, 0.5)$.

A sufficient condition of identifiability is to impose $\tau_{kb}^h > 0$ for all $h = 1, \dots, m_1^{\{kb\}}$. This distribution has very limited interest because it is so unrealistic that it can almost never be used alone. We will see in the next section how to use it in a more efficient way.

3.2 A new block distribution: mixture of two extreme distributions

We proposed to model the distribution of each block by a bi-components mixture between an *independence* distribution and a *maximum dependency* distribution. For block b of component k ,

the block distribution is modeled by:

$$p(\mathbf{x}^{\{kb\}}; \boldsymbol{\theta}_{kb}) = (1 - \rho_{kb})\overset{\circ}{p}(\mathbf{x}^{\{kb\}}; \boldsymbol{\alpha}_{kb}) + \rho_{kb}\overset{\circ}{p}(\mathbf{x}^{\{kb\}}; \boldsymbol{\tau}_{kb}, \boldsymbol{\delta}_{kb}), \quad (5)$$

where $\boldsymbol{\theta}_{kb} = (\rho_{kb}, \boldsymbol{\alpha}_{kb}, \boldsymbol{\tau}_{kb}, \boldsymbol{\delta}_{kb})$ and where ρ_{kb} is the proportion of the maximum dependency distribution in this mixture with $0 \leq \rho_{kb} \leq 1$. The proposed model requires few additional parameters compared with the conditional independence model. In addition, it is easily interpretable as explained in the next paragraph. Note that the limiting case where $\rho_{kb} = 0$ considers that the block follows an independence distribution. In this particular case, the parameters of the maximum dependency distribution are no longer defined.

Under this distribution, the proportion of the maximum dependency distribution reflects the deviation from independence under the assumption that the other allowed distribution is the maximum dependency distribution. The parameter ρ_{kb} gives an indicator of the *inter-variables correlation* of the block. It is not here a pairwise dependency among variables but a dependency between all variables of the block. Furthermore, it stays bounded when the number of variables is larger than two while the Cramer's V is non-upper-bounded in this case. The *intra-variables dependencies* between the variables are defined by $\boldsymbol{\delta}_{kb}$. The strength of these dependencies is explained by $\boldsymbol{\tau}_{kb}$ since it gives the *weight of the over-represented modalities crossing* compared with the independence distribution.

Above, we interpreted the distribution by conditionally independent blocks as a parsimonious version of the log-linear mixture model because it determines the interactions to be modeled for each class. By choosing the proposed distribution for blocks, a second level of parsimony is added. Indeed, among the interactions allowed by this distribution with independent blocks, only those corresponding to the maximum dependency distribution will be modeled. Other interactions are considered as null.

Properties:

- The CCM, stays parsimonious compared with CIM since, for each block with at least two variables, the number of the additional parameters depends only on the number of modalities of the first variable of the block and not on the number of variables in the block. By using

ν_{CIM} defined in Equation (2), the number of parameters of CCM is denoted ν_{CCM} by:

$$\nu_{\text{CCM}} = \nu_{\text{CIM}} + \sum_{\{(k,b)|d^{\{kb\}} > 1\}} m_1^{\{kb\}}. \quad (6)$$

- The proposed distribution is identifiable under the condition that the block is composed by at least three variables ($d^{\{kb\}} > 2$) or that the modality number of the last variable of the block is more than two ($m_2^{\{kb\}} > 2$). This result is demonstrated in Appendix B of Marbac *et al.* (2013). The parameter ρ_{kb} is a new indicator allowing measuring the correlation between variables, not limited to correlation between variable couples. In the case where the identifiability conditions cannot be met, we distinguish two cases. If $d^{\{kb\}} = 1$, then block b contains only one variable, and the proposed model is reduced to model a multinomial distribution, $\rho_{kb} = 0$ and the maximum dependency distribution is not defined. If $d^{\{kb\}} = 2$ and $m_2^{\{kb\}} = 2$ then a new constraint is added. In order to have the most meaningful parameters, the chosen value of ρ_{kb} is the largest value maximizing the log-likelihood. This additional constraint does not falsify the definition of ρ_{kb} as an indicator of the dependency strength between the variables of the same block. Furthermore, this constraint is natural since blocks with the biggest dependencies are wanted. Note that ρ_{kb} seems to be correlated with the Cramer's V. An example is given in Section 3 of Marbac *et al.* (2013).
- Note that the marginal probabilities given class can be straightforwardly deduced from the parameters of the model.

4 Estimation of the parameters

For a fixed model (g, σ) , the parameters must be estimated. Since the proposed distribution CCM has two latent variables (the class membership and the intra-block distribution membership), two algorithms derived from the EM algorithm perform the estimation of the associated continuous parameters. The combinatorial problems arising from the consideration of the discrete parameters are avoided by using a Metropolis-Hastings algorithm.

4.1 Global GEM algorithm

The whole data set consisting of n independent and identically distributed individuals is denoted by $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ where $\mathbf{x}_i \in \mathcal{X}$. The objective is to obtain the maximum log-likelihood estimator $\hat{\boldsymbol{\theta}}$ defined as (g is now implicit in each expression)

$$\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \mathbf{x}, \boldsymbol{\sigma}) \quad \text{with} \quad L(\boldsymbol{\theta}; \mathbf{x}, \boldsymbol{\sigma}) = \sum_{i=1}^n \ln \left(p(\mathbf{x}_i; \boldsymbol{\sigma}, \boldsymbol{\theta}) \right). \quad (7)$$

The search for maximum likelihood estimates for mixture models entails solving equations having no analytical solutions. For the mixture models, the assignments of the individuals into the classes can be considered as missing data. So, the tool generally used is the Expectation-Maximization algorithm (denoted EM algorithm) and its extensions (Dempster *et al.*, 1977; McLachlan & Krishnan, 1997). We denote the unknown indicator vectors of the g clusters by $\mathbf{z} = (z_{ik}; i = 1, \dots, n; k = 1, \dots, g)$ where $z_{ik} = 1$ if \mathbf{x}_i arises from cluster k , $z_{ik} = 0$ otherwise. Thus, the mixture model distribution corresponds to the marginal distribution of the random variable \mathbf{X} obtained from the couple distribution of the random variables (\mathbf{X}, \mathbf{Z}) . In order to maximize the log-likelihood, the EM algorithm uses the complete-data log-likelihood which is defined as

$$L_c(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}, \boldsymbol{\sigma}) = \sum_{i=1}^n \sum_{k=1}^g z_{ik} \ln \left(\pi_k p(\mathbf{x}_i; \boldsymbol{\sigma}_k, \boldsymbol{\theta}_k) \right). \quad (8)$$

The EM algorithm is an iterative algorithm which alternates between two steps: the computation of the complete-data log-likelihood conditional expectation (E step) and its maximization (M step). Many algorithms are derived from the EM algorithm and among them the Generalized EM algorithm (GEM) is of interest to us. It works on the same principle as the EM algorithm, but the maximization step is replaced by a GM step where the proposed parameters increase the expectation of the complete-data log-likelihood according to its previous value without necessarily maximizing it.

We prefer to use the GEM algorithm, since the maximization step in the EM algorithm requires estimating the continuous parameters for too many possible values of the discrete parameters in order to warrant the maximization of the complete-data log-likelihood expectation. Indeed, exhaustive enumeration for estimating the discrete parameters is generally impossible when a block contains variables with many modalities and/or many variables, as detailed below. If $S(a, b)$ is the number of possible surjections from a set of cardinal a into a set of cardinal b , then $\boldsymbol{\delta}_{kb}$ is

defined in the discrete space of dimension $\prod_{j=1}^{d^{\{kb\}}-1} S(m_j^{\{kb\}}, m_{j+1}^{\{kb\}})$. For example, a block with three variables and $m^{\{kb\}} = (5, 4, 3)$ implies 51,840 possibilities for $\boldsymbol{\delta}_{kb}$. Thus, a stochastic approach is proposed in Section 4.2 to overcome this problem. Then, the estimation of the continuous parameters conditionally on the discrete parameters is performed via the classical EM algorithm presented in Section 4.3 since their estimation cannot be obtained in closed form. At iteration (r) , the steps of the global GEM can be written as:

- **E_{global} step:** $z_{ik}^{(r)} = \frac{\pi_k^{(r)} p(\mathbf{x}_i; \boldsymbol{\sigma}_k, \boldsymbol{\theta}_k^{(r)})}{\sum_{k'=1}^g \pi_{k'}^{(r)} p(\mathbf{x}_i; \boldsymbol{\sigma}_{k'}, \boldsymbol{\theta}_{k'}^{(r)})}$,
- **GM_{global} step:** $\pi_k^{(r+1)} = \frac{n_k^{(r)}}{n}$ where $n_k^{(r)} = \sum_{i=1}^n z_{ik}^{(r)}$ and $\forall(k, b)$ $\boldsymbol{\theta}_{kb}^{(r+1)}$ is updated under the constraint that the conditional expectation of complete-data log-likelihood increases (see Sections 4.2 and 4.3).

Initialization of the algorithm: Since this algorithm is performed in an stochastic algorithm used for the model selection (see Section 5) and since this latter has an influence on the GEM initialization, this point will be detailed in Section 5.2.

Stopping criterion: The GEM algorithm is stopped after r_{\max} iterations and we fix $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(r_{\max})}$.

4.2 Details of the GM_{global} step of the GEM

The maximization of the expected complete-data log-likelihood is done by optimizing its terms for each (k, b) . Thus, the determination of $\boldsymbol{\theta}_{kb}^{(r+1)}$ is performed independently of the parameters of the other blocks. A Metropolis-Hastings algorithm (Robert & Casella, 2004) is also performed, for each (k, b) , to avoid the combinatorial problems induced by the detection of the discrete parameters $\boldsymbol{\delta}_{kb}$. It performs a random walk over the discrete parameters space and computes the maximum likelihood estimators of the continuous parameters $(\rho_{kb}, \boldsymbol{\alpha}_{kb}, \boldsymbol{\tau}_{kb})$ associated with them. This stochastic algorithm allows finding the estimator maximizing the expected complete-data log-likelihood of block b for component k :

$$\underset{\boldsymbol{\theta}_{kb}}{\operatorname{argmax}} \sum_{i=1}^n z_{ik}^{(r)} \ln p(\mathbf{x}_i^{\{kb\}}; \boldsymbol{\theta}_{kb}). \quad (9)$$

At each iteration (s) of this Metropolis-Hastings algorithm, a discrete parameter denoted by $\boldsymbol{\delta}_{kb}^{(r, s+\frac{1}{2})}$ is sampled with a uniform distribution in a neighborhood of $\boldsymbol{\delta}_{kb}^{(r, s)}$ denoted as $\Delta(\boldsymbol{\delta}_{kb}^{(r, s)})$.

Then the continuous parameters $(\rho_{kb}^{(r,s+\frac{1}{2})}, \alpha_{kb}^{(r,s+\frac{1}{2})}, \tau_{kb}^{(r,s+\frac{1}{2})})$ are computed, conditionally on the value of $\delta_{kb}^{(r,s+\frac{1}{2})}$, in order to maximize the expected complete-data log-likelihood of block b for component k :

$$\sum_{i=1}^n z_{ik}^{(r)} \ln p(\mathbf{x}_i^{\{kb\}}; \rho_{kb}, \alpha_{kb}, \tau_{kb}, \delta_{kb}^{(r,s+\frac{1}{2})}). \quad (10)$$

The candidate parameters are now denoted by $\theta_{kb}^{(r,s+\frac{1}{2})} = (\rho_{kb}^{(r,s+\frac{1}{2})}, \alpha_{kb}^{(r,s+\frac{1}{2})}, \tau_{kb}^{(r,s+\frac{1}{2})}, \delta_{kb}^{(r,s+\frac{1}{2})})$. The whole block parameters $\theta_{kb}^{(r,s+1)}$ of the next step are then defined as $\theta_{kb}^{(r,s+\frac{1}{2})}$ with the acceptance probability $\mu^{(r,s+1)}$ and $\theta_{kb}^{(r,s)}$ otherwise, where:

$$\mu^{(r,s+1)} = \min \left\{ \frac{\prod_{i=1}^n p(\mathbf{x}_i^{\{kb\}}; \theta_{kb}^{(r,s+\frac{1}{2})})^{z_{ik}^{(r)}} |\Delta(\delta_{kb}^{(r,s+\frac{1}{2})})|}{\prod_{i=1}^n p(\mathbf{x}_i^{\{kb\}}; \theta_{kb}^{(r,s)})^{z_{ik}^{(r)}} |\Delta(\delta_{kb}^{(r,s)})|}, 1 \right\}, \quad (11)$$

$|\Delta(\delta_{kb}^{(r,s)})|$ denoting the cardinal of $\Delta(\delta_{kb}^{(r,s)})$. Thus, at iteration (s) , the algorithm performs the three following steps:

- **Stochastic step on δ_{kb} :** generate $\delta_{kb}^{(r,s+\frac{1}{2})}$ with a uniform distribution among the elements of $\Delta(\delta_{kb}^{(r,s)})$,
- **Maximization step on the continuous parameters (M_θ step):** compute the continuous parameters of $\theta_{kb}^{(r,s+\frac{1}{2})}$ (see Section 4.3),
- **Stochastic step on θ_{kb} :** sample $\theta_{kb}^{(r,s+1)} = \begin{cases} \theta_{kb}^{(r,s+\frac{1}{2})} & \text{with probability } \mu^{(r,s+1)} \\ \theta_{kb}^{(r,s)} & \text{otherwise.} \end{cases}$

Neighborhood $\Delta(\delta_{kb}^{(r,s)})$ is defined as the set of parameters where at most two surjections are different from that of $\delta_{kb}^{(r,s)}$. Figure 3 illustrates this definition.

Initialization of the algorithm: The initialization of the algorithm is done by $\theta_{kb}^{(r+1,0)} = \theta_{kb}^{(r)}$.

Stopping criterion: This algorithm is stopped after a number of iterations s_{\max} . The parameter $\theta_{kb}^{(r+1)} = \theta_{kb}^{(r+1,\tilde{s})}$ is returned with $\tilde{s} = \underset{s}{\operatorname{argmax}} \sum_{i=1}^n z_{ik}^{(r)} \ln p(\mathbf{x}_i^{\{kb\}}; \theta_{kb}^{(r,s)})$. Thus, the proposed initialization ensures an increased likelihood at each iteration of the GEM algorithm.

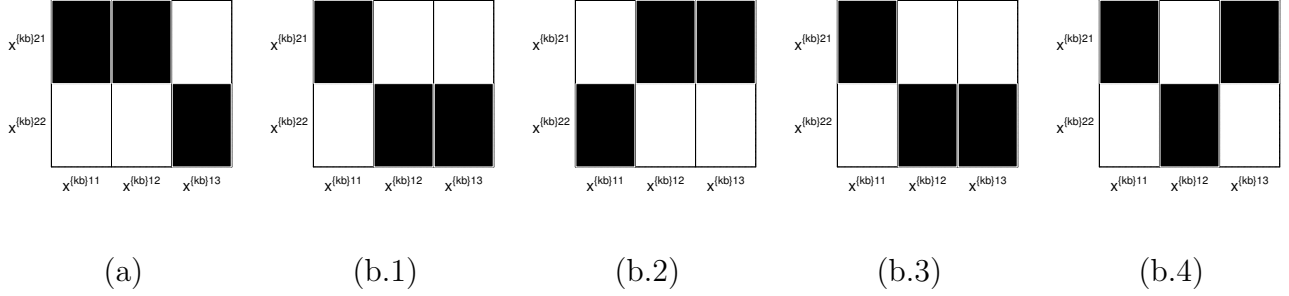


Figure 3: Example of $\Delta(\delta_{kb})$ with $d^{\{kb\}} = 2$ and $\mathbf{m}^{\{kb\}} = (3, 2)$. For row h' and column h , a black cell indicates that $\delta_{kb}^{h2h'} = 1$ and a white cell that $\delta_{kb}^{h2h'} = 0$: (a) δ_{kb} ; (b.1), (b.2), (b.3), (b.4) are the elements of $\Delta(\delta_{kb})$.

Remark: When the space of possible δ_{kb} is small (for example when the block groups a small number of binary variables), an exhaustive approach obtains the same results as the proposed algorithm with less computation time. Thus, the retained approach (exhaustive or stochastic) depends on the number of variables and modalities.

4.3 Details of the M_θ step of the GM_{global} step

As there is a second level of mixing, another EM algorithm can be performed for the continuous parameters $(\rho_{kb}, \alpha_{kb}, \tau_{kb})$ estimation by introducing other unknown vectors corresponding to the indicator of the blocks distributions conditionally on \mathbf{z} . These vectors are denoted by $\mathbf{y} = (\mathbf{y}^{\{kb\}}; k = 1, \dots, g; b = 1, \dots, B_k)$ with $\mathbf{y}^{\{kb\}} = (y_1^{\{kb\}}, \dots, y_n^{\{kb\}})$ where $y_i^{\{kb\}} = 1$ if $\mathbf{x}_i^{\{kb\}}$ arises from the *maximum dependency* distribution for block b of cluster k and $y_i^{\{kb\}} = 0$ if $\mathbf{x}_i^{\{kb\}}$ arises from the *independence* distribution for block b of cluster k . The whole mixture model distribution corresponds to the marginal distribution of the random variable \mathbf{X} obtained from the triplet distribution of the random variables $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$. Since the blocks are independent conditionally on \mathbf{Z} , the *full* complete-data log-likelihood (both in \mathbf{Y} and \mathbf{Z}) is defined as:

$$L_c^{\text{full}}(\theta; \mathbf{x}, \mathbf{y}, \mathbf{z}, \sigma) = \sum_{i=1}^n \sum_{k=1}^g z_{ik} \left(\ln \pi_k + \sum_{b=1}^{B_k} ((1 - y_i^{\{kb\}}) \ln(1 - \rho_{kb}) + (1 - y_i^{\{kb\}}) \ln \dot{p}(\mathbf{x}_i^{\{kb\}}; \alpha_{kb}) + y_i^{\{kb\}} \ln \rho_{kb} + y_i^{\{kb\}} \ln \dot{p}(\mathbf{x}_i^{\{kb\}}; \tau_{kb}, \delta_{kb})) \right). \quad (12)$$

At iteration (t) , the local EM algorithm estimates the continuous parameters of block b , with fixed values of $\mathbf{z}^{(r)}$ and $\delta_{kb}^{(r, s + \frac{1}{2})}$, by the following two steps:

- **E_{local} step:** $y_i^{\{kb\}(r, s+\frac{1}{2}, t)} = \frac{\rho_{kb}^{(r, s+\frac{1}{2}, t)} \dot{p}(x_i^{\{kb\}}; \tau_{kb}^{(r, s+\frac{1}{2}, t)}, \delta_{kb}^{(r, s+\frac{1}{2}, t)})}{p(x_i^{\{kb\}}; \rho_{kb}^{(r, s+\frac{1}{2}, t)}, \alpha_{kb}^{(r, s+\frac{1}{2}, t)}, \tau_{kb}^{(r, s+\frac{1}{2}, t)}, \delta_{kb}^{(r, s+\frac{1}{2}, t)})},$
- **M_{local} step:** $\rho_{kb}^{(r, s+\frac{1}{2}, t+1)} = \frac{n_{kb}^{(r, s+\frac{1}{2}, t)}}{n_k^{(r)}}, \tau_{kb}^{(r, s+\frac{1}{2}, t+1)} = \frac{\sum_{i=1}^n z_{ik}^{(r)} y_i^{\{kb\}(r, s+\frac{1}{2}, t)} x_i^{\{kb\}1h}}{n_{kb}^{(r, s+\frac{1}{2}, t)}},$
 $\alpha_{kb}^{(r, s+\frac{1}{2}, t+1)} = \frac{\sum_{i=1}^n z_{ik}^{(r)} (1 - y_i^{\{kb\}(r, s+\frac{1}{2}, t)}) x_i^{\{kb\}jh}}{n_k^{(r)} - n_{kb}^{(r, s+\frac{1}{2}, t)}},$ where $n_{kb}^{(r, s+\frac{1}{2}, t)} = \sum_{i=1}^n z_{ik}^{(r)} y_i^{\{kb\}(r, s+\frac{1}{2}, t)}.$

Conjecture: During our numerous experiments, we empirically noticed that the log-likelihood function of the mixture between the independence and the maximum dependency distributions had a unique optimum. We conjecture that this function has indeed a unique maximum.

Initialization of the algorithm: The previous conjecture allows to perform only one initialization of the EM algorithm fixed to: $(\rho_{kb}^{(r, s+\frac{1}{2}, 0)}, \alpha_{kb}^{(r, s+\frac{1}{2}, 0)}, \tau_{kb}^{(r, s+\frac{1}{2}, 0)}) = (\rho_{kb}^{(r, s)}, \alpha_{kb}^{(r, s)}, \tau_{kb}^{(r, s)}).$

Stopping criterion: This algorithm is stopped after a number of iterations denoted by t_{\max} and returns the value of block parameters $\theta_{kb}^{(r, s+\frac{1}{2})}$ defined as $\theta_{kb}^{(r, s+\frac{1}{2})} = \theta_{kb}^{(r, s+\frac{1}{2}, t_{\max})}.$

Remark: In the specific case where δ_{kb} are known for each (k, b) , the estimation of all the continuous parameters could be performed by a unique EM algorithm where, at iteration (r) , the E step would compute both $\mathbf{z}^{(r)}$ and $\mathbf{y}^{(r)}$ while the M step would estimate all the parameters maximizing the expectation of the *full* complete-data log-likelihood.

5 Model selection

5.1 Gibbs algorithm for exploring the space of models

Since the number of components g determines the dimension of σ , the model construction is done in two steps. Firstly, the selection of the number of components and, secondly, the determination of the variable repartition per blocks for each component. In a Bayesian context, the best model $(\hat{g}, \hat{\sigma})$ is defined as (Robert, 2005):

$$(\hat{g}, \hat{\sigma}) = \underset{g, \sigma}{\operatorname{argmax}} p(g, \sigma | \mathbf{x}). \quad (13)$$

Thus, by considering that $p(g) = \frac{1}{g_{\max}}$ if $g \leq g_{\max}$ and 0 otherwise, where g_{\max} is the maximum number of classes allowed by the user, and by assuming that $p(\boldsymbol{\sigma}|g)$ follows a uniform distribution, the best model is also defined as:

$$(\hat{g}, \hat{\boldsymbol{\sigma}}) = \operatorname{argmax}_g \left[\operatorname{argmax}_{\boldsymbol{\sigma}} p(\mathbf{x}|g, \boldsymbol{\sigma}) \right]. \quad (14)$$

To obtain $(\hat{g}, \hat{\boldsymbol{\sigma}})$, a Gibbs algorithm is used for estimating $\operatorname{argmax}_{\boldsymbol{\sigma}} p(\mathbf{x}|g, \boldsymbol{\sigma})$, for each value of $g \in \{1, \dots, g_{\max}\}$. Indeed, this method limits the combinatorial problem involved by the detection of the block structure of variables. A reversible jump method could be used (Richardson & Green, 1997), but this approach is rarely performed with mixed parameters (continuous and discrete). Indeed, in such a case, it is difficult to define a mapping between the parameters space of two models. So, we propose to use an easier Gibbs sampler-type having $p(\boldsymbol{\sigma}|\mathbf{x}, g)$ as stationary distribution. It alternates between two steps: the generation of a stochastic neighborhood $\Sigma^{[q]}$ conditionally on the current model $\boldsymbol{\sigma}^{[q]}$ by a proposal distribution and the generation of a new pattern $\boldsymbol{\sigma}^{[q+1]}$ included in $\Sigma^{[q]}$ with a probability proportional to its posterior probability. At iteration $[q]$, it is written as:

- **Neighborhood step:** generate a stochastic neighborhood $\Sigma^{[q]}$ by a proposal distribution given below conditionally on the current model $\boldsymbol{\sigma}^{[q]}$,

- **Pattern step:** $\boldsymbol{\sigma}^{[q+1]} \sim p(\boldsymbol{\sigma}|\mathbf{x}, g, \Sigma^{[q]})$ with $p(\boldsymbol{\sigma}|\mathbf{x}, g, \Sigma^{[q]}) = \begin{cases} \frac{p(\mathbf{x}|g, \boldsymbol{\sigma})}{\sum_{\boldsymbol{\sigma}' \in \Sigma^{[q]}} p(\mathbf{x}|g, \boldsymbol{\sigma}')} & \text{if } \boldsymbol{\sigma} \in \Sigma^{[q]} \\ 0 & \text{otherwise.} \end{cases}$

A possible deterministic neighborhood of $\boldsymbol{\sigma}^{[q]}$ could be defined as the set of models where, at most one variable is affected, for one component, in another block (possibly creating a new block): $\{\boldsymbol{\sigma} : \exists!(k, b, j) \ j \in \boldsymbol{\sigma}_{kb}^{[q]} \text{ and } j \notin \boldsymbol{\sigma}_{kb} \} \cup \{\boldsymbol{\sigma}^{[q]}\}$. However, as this deterministic neighborhood can be very large, our proposal distribution allows reducing it to a stochastic neighborhood $\Sigma^{[q]}$ by reducing the number of (k, b) where $\boldsymbol{\sigma}_{kb}$ could be different to $\boldsymbol{\sigma}_{kb}^{[q]}$. Thus, one component $k^{[q]}$ is randomly sampled in $\{1, \dots, g\}$ then one block $b_{\text{from}}^{[q]}$ is randomly sampled in $\{1, \dots, B_{k^{[q]}}^{[q]}\}$. Another block $b^{[q]}$ is randomly sampled in $\{1, \dots, B_{k^{[q]}}^{[q]}\} \setminus b_{\text{from}}^{[q]}$ and the set $b_{\text{to}}^{[q]} = \{b^{[q]}, B_{k^{[q]}}^{[q]} + 1\}$ is built. The stochastic neighborhood $\Sigma^{[q]}$ is then defined as:

$$\Sigma^{[q]} = \left\{ \boldsymbol{\sigma} : \exists!(k, b, j) \ j \in \boldsymbol{\sigma}_{kb}^{[q]}, \ j \notin \boldsymbol{\sigma}_{kb} \text{ and } j \in \boldsymbol{\sigma}_{kb'} \text{ with } k = k^{[q]}, \ b = b_{\text{from}}^{[q]}, \ b' \in b_{\text{to}}^{[q]} \right\} \cup \left\{ \boldsymbol{\sigma}^{[q]} \right\}. \quad (15)$$

We denote the elements of $\Sigma^{[q]}$ as $\boldsymbol{\sigma}^{[q+\varepsilon(e)]}$ where $\varepsilon(e) = \frac{e}{|\Sigma^{[q]}|+1}$ and $e = 1, \dots, |\Sigma^{[q]}|$. Figure 4 shows an illustration of this definition.

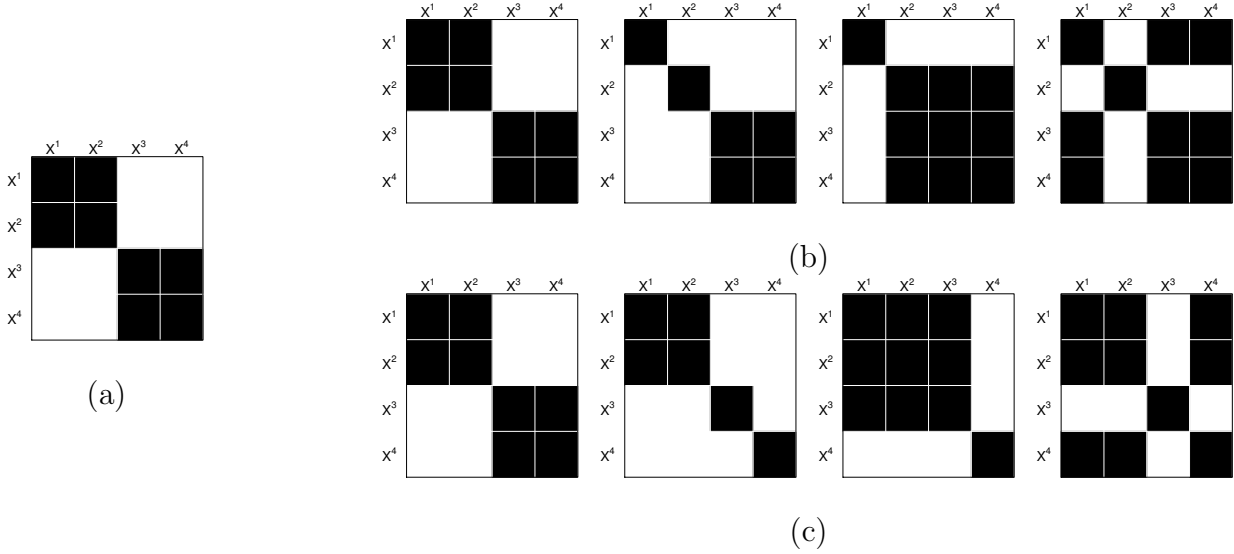


Figure 4: Example of the support of $\Sigma^{[q]}$ in a case of four variables. If the variable of row j and the variable of column j' are in the same block then cell (j, j') is painted in black. This cell is painted in white otherwise. (a) Graphical representation of $\sigma_k^{[q]} = (\{1, 2\}, \{3, 4\})$; (b) Elements of $\Sigma^{[q]}$ if $b_{from}^{[q]} = 1$; (c) Elements of $\Sigma^{[q]}$ if $b_{from}^{[q]} = 2$.

At the generation pattern step, the previous algorithm needs the value of $p(\mathbf{x}|g, \sigma) \forall \sigma \in \Sigma^{[q]}$. By using the BIC approximation (Schwarz, 1978; Lebarbier & Mary-Huard, 2006), this probability is approximated by:

$$\ln p(\mathbf{x}|g, \sigma) \simeq L(\hat{\theta}; \mathbf{x}, g, \sigma) - \frac{\nu_{\text{CCM}}}{2} \log(n), \quad (16)$$

$\hat{\theta}$ being the maximum likelihood estimator obtained by the GEM algorithm previously described in Section 4. Thus, at the iteration $[q]$, for each $e = 1, \dots, |\Sigma^{[q]}|$, estimator $\hat{\theta}^{[q+\varepsilon(e)]}$ associated to element $\sigma^{[q+\varepsilon(e)]}$ is computed by the GEM algorithm.

Initialization: Whatever the initial value selected for $\sigma^{[0]}$, the algorithm converges to the same value of σ . However, this convergence can be very slow when the initialization is poor. Since blocks consist of the most correlated variables, a Hierarchical Ascendant Classification (HAC) is used on the matrix of Cramer's V distances on the variable couples. The partition produced by the HAC minimizing the block number without blocks consisting of more than four variables is chosen for each $\sigma_k^{[0]}$. For the initialization, the variables number of a block is limited to four, because very few blocks having more than four variables were exhibited in the course of our experiments. Obviously, the Gibbs algorithm can then violate this initial constraint if necessary.

Stopping criterion: The algorithm is stopped when q_{\max} successive iterations have not discovered a better model.

5.2 Consequences of the Gibbs algorithm on the GEM algorithm

Initialization of the GEM algorithm: At iteration $[q]$ of the Gibbs algorithm, the GEM algorithm estimates $\hat{\boldsymbol{\theta}}^{[q+\varepsilon(e)]}$ associated to model $\boldsymbol{\sigma}^{[q+\varepsilon(e)]}$ for $e = 1, \dots, |\Sigma^{[q]}|$. Since these models are close to $\boldsymbol{\sigma}^{[q]}$, their maximum likelihood estimators should be closed to $\hat{\boldsymbol{\theta}}^{[q]}$. The GEM algorithm initialization is also done by the value of $\hat{\boldsymbol{\theta}}^{[q]}$ for the not modified blocks. Thus, $\boldsymbol{\theta}_{kb}^{[q+\varepsilon(e)](0)} = \hat{\boldsymbol{\theta}}_{kb}^{[q]}$ if the blocks are not modified ($\boldsymbol{\sigma}_{kb}^{[q+\varepsilon(e)]} = \boldsymbol{\sigma}_{kb}^{[q]}$). For the other blocks, the continuous parameters are randomly sampled. For those blocks, in order to avoid the combinatorial problems, we use a sequential method to initialize $\boldsymbol{\delta}_{kb}^{[q+\varepsilon(e)](0)}$: the surjections from $\mathbf{x}^{\{kb\}1}$ to $\mathbf{x}^{\{kb\}j}$ are sampled, according to \mathbf{x} and to the continuous parameters previously sampled ($\rho_{kb}^{[q+\varepsilon(e)](0)}, \boldsymbol{\alpha}_{kb}^{[q+\varepsilon(e)](0)}, \boldsymbol{\tau}_{kb}^{[q+\varepsilon(e)](0)}$), for each $j = 2, \dots, d^{\{kb\}}$ as follows:

$$\boldsymbol{\delta}_{kb}^{j[q+\varepsilon(e)](0)} \propto \prod_{i=1}^n p(x_i^{\{kb\}1}, x_i^{\{kb\}j}; \rho_{kb}^{[q+\varepsilon(e)](0)}, \boldsymbol{\alpha}_{kb}^{1[q+\varepsilon(e)](0)}, \boldsymbol{\alpha}_{kb}^{j[q+\varepsilon(e)](0)}, \boldsymbol{\tau}_{kb}^{[q+\varepsilon(e)](0)}, \boldsymbol{\delta}_{kb}^{j} z_{ik}^{[q]}), \quad (17)$$

where $\boldsymbol{\delta}_{kb}^{j[q+\varepsilon(e)]} = (\boldsymbol{\delta}_{kb}^{hj[q+\varepsilon(e)]}; h = 1, \dots, m_1^{\{kb\}})$ and where $z_{ik}^{[q]} = E[Z_{ik} | \mathbf{x}_i, \boldsymbol{\theta}^{[q]}]$.

Remark about r_{\max} : As said in Section 4.1, the algorithm is stopped after a fixed number of iterations r_{\max} . If the algorithm is stopped before its convergence, the proposed initialization limits the problems. Indeed, if the model has a high *a posteriori* probability, it will stay in neighborhood $\Sigma^{[q]}$ during some successive iterations, so its log-likelihood will increase.

6 Simulations

Table 1 presents the adjustment parameters values used for all the simulations.

Algorithms	Gibbs	GEM	Metropolis-Hastings	EM
Criteria	$q_{\max} = 20 \times d$	$r_{\max} = 10$	$s_{\max} = 1$	$t_{\max} = 5$

Table 1: Values of the different stopping criteria.

As these algorithms are interlocked, the iterations number of the most internal algorithms are small. Since the number of possible models increases with d , we propose to set: $q_{\max} = 20 \times d$. When the best model is selected by the Gibbs algorithm, this latter will stay in this model during many iterations so the Metropolis-Hastings and the EM algorithm are performed many times. Thus, it is not necessary to have a large number of iterations as stopping criterion.

6.1 Study of the algorithm for the δ_{kb} estimation

In this section, we illustrate the performance of the Metropolis-Hastings algorithm used for the δ_{kb} estimation (see Section 4.2) and the relevance of its initialization (see Equation (17)). Since this algorithm is interlocked in the Gibbs and in the GEM algorithm, we need it to converge quickly. The following simulations show that the algorithm stays relevant up to six modalities per variable and up to six variables per block. These conditions hold in most situations.

Samples of size 200 described by variables having the same number of modalities are generated by a mixture between an independence distribution and a maximum dependency distribution. The parameter estimation is also performed by the Metropolis-Hastings algorithm, described in Section 4.2, since only one class is generated. The discrete parameters initializations are performed according to Equation (17) with $z_{i1} = 1$ for $i = 1, \dots, 200$.

Figure 5 shows the box-plots of the iterations number needed by the Metropolis-Hastings algorithm for finding the true links between modalities maximizing the likelihood². According to these simulations, one observes that the results of this algorithm are good as a result of its initialization which allows significantly reducing the number of iterations needed in order to find the maximum likelihood estimators.

6.2 Study of the algorithm for model selection

In order to illustrate the efficiency of the algorithm for the model selection (and also the included estimation process), we want to study the evolution of the Kullback-Leibler divergence according to the number of variables and to the size of the data set. Thus, 100 samples are generated for many situations according to the CCM with two components. Note that the parameter u is introduced

²In fact, the algorithm is stopped as soon as it finds a discrete estimator involving a likelihood higher than or equal to the likelihood obtained with the true discrete parameters used for the simulation.

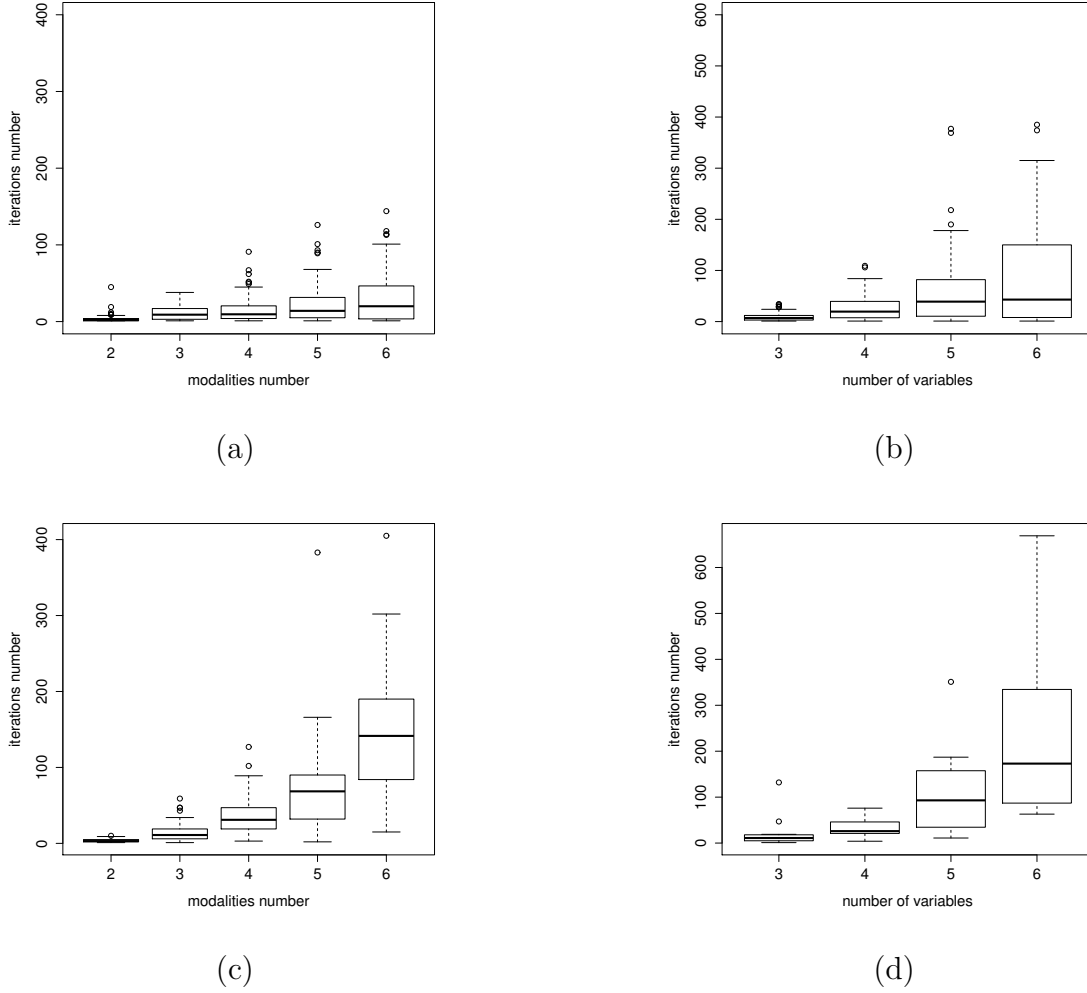


Figure 5: Box-plots of the number of iterations required by the Metropolis-Hastings algorithm for finding the best links between modalities, according to the number of modalities when datasets are simulated with a proportion of maximum dependency distribution equal to 0.5. (a) Three variables with the proposed initialization; (b) Three modalities per variable with the proposed initialization; (c) Three variables with a random initialization; (d) Three modalities per variable with a random initialization.

for controlling the overlapping of classes: when it is close to one their overlapping (Bayes error) is close to one. These parameters set the error rate to 0.10 for each studied situation:

$$\sigma_{kb} = (d/b, 1 + d/b) \quad \rho_{kb} = 0.6(1 - u) \quad \tau_{kb} = (0.60, 0.20, 0.20),$$

$$\delta_{1b}^{h2h'} = 1 \text{ iff } h = h' \quad \delta_{1b}^{122} = \delta_{1b}^{223} = \delta_{1b}^{321} = 1 \quad \alpha_{1b}^j = (0.20, 0.20, 0.60),$$

$$\alpha_{2b}^1 = \alpha_{1b}^1(1 - u) + (0.075, 0.850, 0.075)u \quad \text{and} \quad \alpha_{2b}^2 = \alpha_{1b}^2(1 - u) + (0.850, 0.075, 0.075)u.$$

Table 2 shows the mean and the standard deviation of the Kullback-Leibler divergence between the parameters used for the dataset generation and the estimated parameters according to the number of variables. When n increases, the Kullback-Leibler divergence converges to zero. It confirms the good behavior of the proposed algorithm.

$d \setminus n$	100	200	400	800
4	0.77 (1.34)	0.26 (0.26)	0.15 (0.05)	0.12 (0.05)
6	1.22 (1.77)	0.27 (0.14)	0.09 (0.07)	0.05 (0.05)
8	1.72 (2.50)	0.41 (0.20)	0.09 (0.05)	0.05 (0.03)
10	1.73 (4.06)	0.52 (0.14)	0.10 (0.03)	0.04 (0.03)

Table 2: **Mean** (*standard deviation*) of the Kullback-Leibler divergence.

7 Application

7.1 Dentistry clustering

The Handelman’s dentistry data (Handelman *et al.* , 1986) display the evaluation of 3869 dental x-rays (sound or carious) performed by five dentist and possibly showing incipient caries. This data set has been clustered by several models in the past. It is suggested that there are two main classes: the sound teeth and the carious ones.

According to the BIC criterion, data are split into three classes by CIM. Furthermore, dependencies are observed between the variables into classes since the Cramer’s V computed per class is not close to zero. Thus, Espeland & Handelman (1989) apply a log-linear mixture model to fit the data. The authors set the model, by adding some assumptions to better fit the data. More precisely, they consider a mixture with four components. The first two components take into account the interactions between dentists 3 and 4. The last two components are specific since they allow only one modality interaction, when all the diagnoses are respectively carious and sound. Note that these assumptions are required by the above authors due to their realistic nature. Indeed, this model fits the data better than CIM. On the other hand, its interpretation needs the analysis of four classes.

As the last two classes seem artificial, Qu *et al.* (1996) prefer to use the random effects models in a latent class analysis with two classes. They assume that the conditional dependencies can be modeled by a single continuous latent variable varying among the individuals. According to the authors, one class represents the sound teeth and the other represents the carious ones, while the random effect represents all the patient-specific unrecorded characteristics of the x-ray images. Their model does not need the two additional artificial classes. Thus their interpretation is easier.

We now display the results of the proposed model CCM estimated with the R package `Clustericat` (the code is presented in Appendix A). The BIC criterion selects two classes with a value of -7473. It claims that CMM better fits the data than the model of Qu *et al.* (1996) since their BIC criterion value is -7487. The BIC criterion values for CIM and CMM are displayed in Table 3. We indicate the computing time (in seconds), obtained with an Intel Core i5-3320M processor, to estimate CMM where 20 MCMC chains were started with a stopping rule $q_{\max} = 100$ while CIM needs less than 0.1 sec with the R package `RMixmod` (Lebrete *et al.* , 2012).

g		1	2	3	4
CIM	BIC	-8766	-7511	-7481	-7503
CCM	BIC	-7743	-7473	-7481	-7503
	time (sec)	1.7	4.9	6.1	7.7

Table 3: BIC criterion values for the CIM and the CMM according to different numbers of classes for the dentistry data set. For each model, the best results according to the BIC criterion are in bold. The computing time in seconds is indicated for CMM where 20 MCMC chains were started with a stopping rule $q_{\max} = 500$.

We note that CMM obtains better values for the BIC criterion than CIM when $g = 1, 2$. When the number of classes is larger ($g \geq 3$) the best model of CMM assumes the conditional independence between variables.

The BIC criterion selects two classes for CMM and this is coherent with a clustering of the teeth between the sound and the carious ones. Furthermore, the two main characteristics of the model fixed by Espeland & Handelman (1989) are automatically detected by the model: importance of the two modality crossings where all the dentists have the same diagnosis and a dependency between the diagnosis of dentists 3 and 4. Thus, the estimated model is coherent with the imposed

model of Espeland & Handelman (1989) while no information was given *a priori*.

The fitted model can be interpreted as:

- the majority class ($\pi_1 = 0.86$) mainly gathers the sound teeth. There is a strong dependency between the five dentists ($\sigma_1 = (\{1, 2, 3, 4, 5\})$ and $\rho_{11} = 0.35$). The dependency structure of the maximum dependency distribution indicates an over-contribution of both modality interactions where the five dentists have the same diagnosis, especially when they claim that the teeth are sound ($\tau_{11}^{\text{all-sound}} = 0.93$ and $\tau_{11}^{\text{all-carious}} = 0.07$). This class could be interpreted as grouping the teeth whose diagnosis is obvious.
- the minority class ($\pi_2 = 0.14$) groups principally the carious teeth. There is a dependency between dentists 3 and 4 while the diagnosis of the other ones are independent given the class ($\sigma_2 = (\{3, 4\}, \{1, 2, 5\})$, $\rho_{21} = 0.31$ and $\rho_{22} = 0$). This class could be interpreted as grouping the teeth whose diagnosis is more complex.

Figure 6 helps the interpretation of the clusters for the CCM with two components (best model according to the BIC criterion). On ordinates, the estimated classes are represented with respect to their proportions in decreasing order. Their corresponding area depends on their proportion. The cumulated proportions are indicated on the left side. On abscissae, three indications are given. The first one is the inter-variables correlations (ρ_{kb}) for all the blocks of the class ordered by their strength of correlation (in decreasing order). The second one is the intra-variables correlations (τ_{kb}) for each block drawn according to their strength dependencies (in decreasing order). The third is the variables repartition per block. A black cell indicates that the variable is assigned to the block and a white cell indicates that, conditionally on this class, the variable is independent of the variables of this block. For example, this figure shows that the first class has a proportion of 0.86 and that all the variables are assigned into the same block.

7.2 Calves clustering

In this section, the results obtained by the CCM are compared to those obtained for the CIM by the RMixmod software (Lebrete *et al.*, 2012). The “Genes Diffusion” company has collected information from French breeders in order to cluster calves. The 4270 studied calves are described by nine variables related to behavior (aptitude for sucking *Apt*, behavior of the mother just before

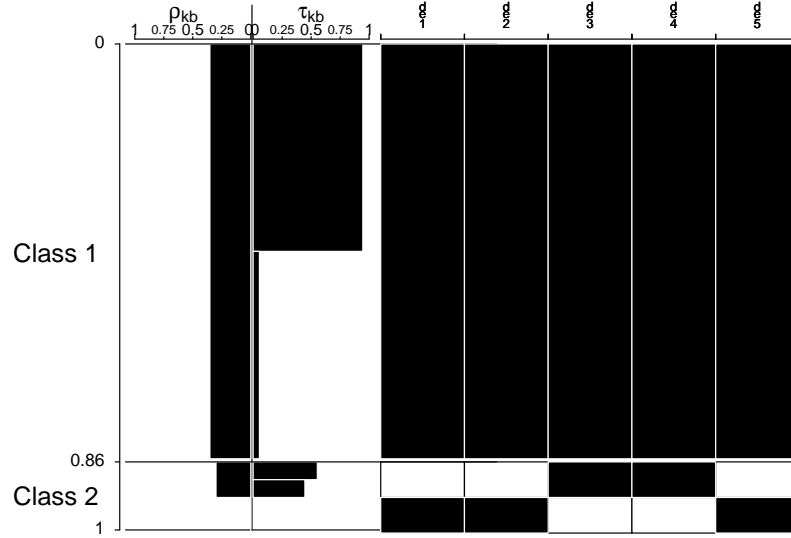


Figure 6: Summary of the best CCM according to BIC for the dentists data set.

the calving *Iso*) and health (treatment against omphalitis *TOC*, respiratory disease *TRC* and diarrhea *TDC*, umbilicus disinfection *Dis*, umbilicus emptying *Emp*, mother preventive treatment against respiratory disease *TRM* and diarrhea *TDM*).

Table 4 displays the BIC criterion values and the number of parameters for the CIM and CCM models. Furthermore, the computing time in minutes (obtained with an Intel Core i5-3320M processor) to estimate CCM by starting 20 MCMC chains with a stopping criterion of $q_{\max} = 180$ while CIM needs 3 seconds with the R package RMixmod (Lebret *et al.*, 2012).

	g	1	2	3	4	5	6	7	8
CIM	BIC	-28589	-26859	-26526	-26333	-26238	-26235	-26226	-26185
	ν_{CIM}	17	35	53	71	89	107	125	143
CCM	BIC	-26653	-26289	-26173	-26038	-26025	-26059	-26045	-26058
	ν_{CCM}	24	48	80	89	112	131	148	163
	time (min)	0.97	3.32	6.16	6.56	10.03	11.76	12.31	14.92

Table 4: Results for the CIM and the CMM according to different class numbers. For both models, first row corresponds to the BIC criterion values and the second row indicates the continuous parameter number. For each model, the best results according to the BIC criterion are in bold. The computing time for the CCM estimation is given in minutes.

For the CIM, the BIC criterion selects a high number of classes, since it selected eight classes. The interpretation of the clusters is also difficult and we can assume that the estimator's quality is very poor. Figure 7 helps the interpretation for the CCM with five components (best model according to the BIC criterion). Its interpretation is the same as the interpretation of Figure 6. For example, this figure shows that the first class has a proportion of 0.29 and that it is composed of four blocks. The most correlated block of the first class has $\rho_{kb} \simeq 0.80$ and the strength of the biggest modalities link is also close to 0.85. This block consists of the variables *TDC* and *TRM*. Here is now a possible interpretation of Class 1 (note that the others classes are also meaningful;

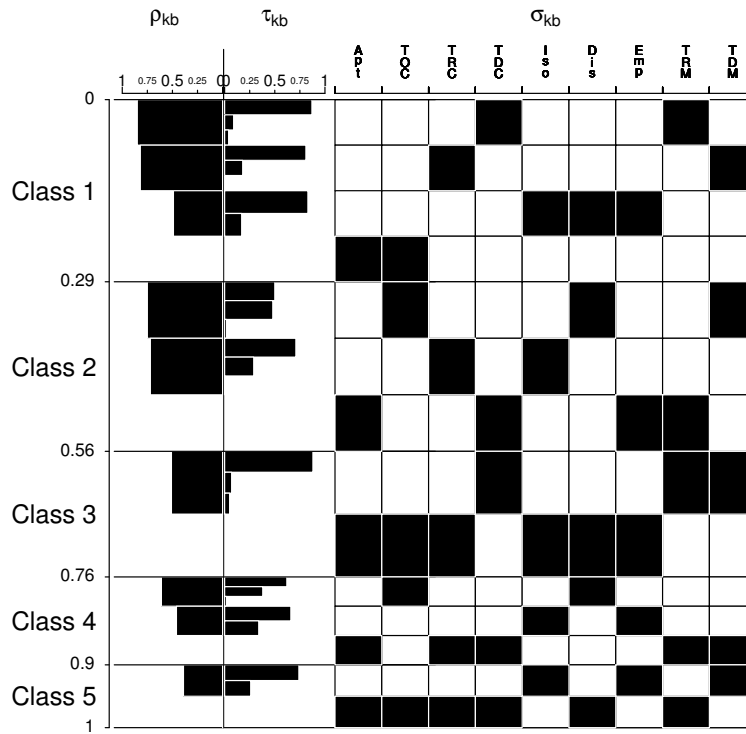


Figure 7: Summary of the best CCM according to BIC for the calves data set.

see details in Marbac *et al.* (2013)):

- **General:** This class has a proportion equal to 0.29 and consists of three blocks of dependency and one block of independence.
- **Block 1:** There is a strong correlation (ρ_{11}) between the variables diarrhea treatment of the calf and mother preventive treatment against respiratory disease, especially between the

modality no treatment against the calf diarrhea and the absence of preventive treatment against respiratory disease of its mother (τ_{11} and δ_{11}).

- **Block 2:** There is a strong correlation (ρ_{12}) between the variables treatment against respiratory illness of the calf and mother preventive treatment against diarrhea, especially between the modality preventive treatment against respiratory illness of the calf and the presence of diarrhea preventive treatment of its mother (τ_{12} and δ_{12}).
- **Block 3:** There exists another strong link between the behavior of the mother, the emptying of the umbilical and its disinfection (τ_{13} and δ_{13}).
- **Block 4:** This block is characterized by absence of preventive treatment against omphalitis and having 50% of the calves infected by this illness (α_{14}).

8 Conclusion

By using the block extension of the latent class model, a new mixture model is proposed for clustering categorical data by taking into account the intra-class correlation. The block distribution is defined as a mixture between an independent distribution and a maximum dependency distribution. This specific distribution, which remains parsimonious, is compared to the full latent class model and allows different levels of interpretation. The blocks of variables detect the conditional dependencies between variables while their strengths are reflected by the proportions of maximum dependency distribution. The parameters of the block distribution reflect the links and the strength between modalities.

The parameter estimation and the model selection are simultaneously performed via a Gibbs sample-type algorithm. It allows reducing the combinatorial problems of the block structure detection and the links between modalities search for the estimation of the maximum dependency distribution. The results are good when the number of modalities is small for each variable. For more than six modalities, the detection of other links encounters some persistent difficulties. So the algorithm can be slow in this case. The proposed approach to estimate the block structure is not adapted for data sets with many variables. A deterministic but sub-optimal solution could be used to perform a forward algorithm.

The R package *Clustericat* allows clustering categorical data sets by using CMM. This package is available on Rforge at the following url https://r-forge.r-project.org/R/?group_id=1803.

The proposed model can be easily extended to the case of ordinal data. For this purpose, some additional constraints on the dependency structure of each distribution of maximum dependency need to be added.

Acknowledgments: The author thank the Associate Editor and the three anonymous referee for their useful comments and references. The authors are grateful to the Genes Diffusion corporation for the provision of the data set and especially its members: Amélie Vallée, Julie Hamon and Claude Grenier. We are grateful to Parmeet Bhatia, Modal Team engineer, and Stéphane Chrétien for their precious assistance. This work was financed by DGA and Inria.

References

- Agresti, A. 2002. *Categorical data analysis*. Vol. 359. John Wiley and Sons.
- Allman, E.S., Matias, C., & Rhodes, J.A. 2009. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, **37**(6A), 3099–3132.
- Banfield, J.D., & Raftery, A.E. 1993. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 803–821.
- Biernacki, C., Celeux, G., Govaert, G., & Langrognet, F. 2006. Model-based cluster and discriminant analysis with the MIXMOD software. *Computational Statistics & Data Analysis*, **51**(2), 587–600.
- Bock, H.H. 1986. Loglinear models and entropy clustering methods for qualitative data. *Classification as a tool of research*. North Holland, Amsterdam, 19–26.
- Celeux, G., & Govaert, G. 1991. Clustering criteria for discrete data and latent class models. *Journal of classification*, **8**(2), 157–176.
- Celeux, G., & Govaert, G. 1995. Gaussian parsimonious clustering models. *Pattern Recognition*, **28**(5), 781–793.

- Chavent, M., Kuentz, V., & Saracco, J. 2010. A partitioning method for the clustering of categorical variables. *Pages 91–99 of: Classification as a Tool for Research*. Springer.
- Cheng, J., & Greiner, R. 1999. Comparing Bayesian network classifiers. *Pages 101–108 of: Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc.
- Chow, C., & Liu, C. 1968. Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, **14**(3), 462–467.
- Dempster, A.P., Laird, N.M., & Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.
- Espeland, M.A., & Handelman, S.L. 1989. Using Latent Class Models to Characterize and Assess Relative Error in Discrete Measurements. *Biometrics*, **45**(2), pp. 587–599.
- Formann, A.K. 1992. Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association*, **87**(418), 476–486.
- Friedman, N., Geiger, D., & Goldszmidt, M. 1997. Bayesian network classifiers. *Machine learning*, **29**(2), 131–163.
- Gollini, I., & Murphy, T.B. 2013. Mixture of latent trait analyzers for model-based clustering of categorical data. *Statistics and Computing*, 1–20.
- Goodman, L.A. 1974. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, **61**(2), 215–231.
- Govaert, G. 2010. *Data analysis*. Vol. 136. Wiley. com.
- Govaert, G., & Nadif, M. 2003. Clustering with block mixture models. *Pattern Recognition*, **36**(2), 463–473.
- Guinot, C., Latreille, J., Malvy, D., Preziosi, P., Galan, P., Herberg, S., & Tenenhaus, M. 2001. Use of multiple correspondence analysis and cluster analysis to study dietary behaviour: food consumption questionnaire in the SU. VI. MAX. cohort. *European journal of epidemiology*, **17**(6), 505–516.

- Hagenaars, J.A. 1988. Latent structure models with direct effects between indicators local dependence models. *Sociological Methods & Research*, **16**(3), 379–405.
- Hand, D.J., & Yu, K. 2001. Idiot’s BayesNot So Stupid after All? *International Statistical Review*, **69**(3), 385–398.
- Handelman, S.L., Leverett, D.H., Espeland, M.A., & Curzon, J.A. 1986. Clinical radiographic evaluation of sealed carious and sound tooth surfaces. *The Journal of the American Dental Association*, **113**(5), 751–754.
- Harper, D. 1972. Local dependence latent structure models. *Psychometrika*, **37**(1), 53–59.
- Huang, J.Z., Ng, M.K., Rong, H., & Li, Z. 2005. Automated variable weighting in k-means type clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **27**(5), 657–668.
- Hunt, L., & Jorgensen, M. 1999. Theory & Methods: Mixture model clustering using the MULTIMIX program. *Australian & New Zealand Journal of Statistics*, **41**(2), 154–171.
- Jajuga, K., Sokołowski, A., & Bock, H.H. 2002. *Classification, clustering and data analysis: recent advances and applications*. Springer Verlag.
- Jorgensen, M., & Hunt, L. 1996. Mixture model clustering of data sets with categorical and continuous variables. *Pages 375–384 of: Proceedings of the Conference ISIS*, vol. 96.
- Lebarbier, E., & Mary-Huard, T. 2006. Une introduction au critre BIC : fondements thoriques et interpretation. *Journal de la SFdS*, **147**(1), 39–57.
- Lebret, R., Iovleff, S., Langrognet, F., Biernacki, C., Celeux, G., & Govaert, G. 2012. Rmixmod: The R Package of the Model-Based Unsupervised, Supervised and Semi-Supervised Classification Mixmod Library. *preprint submitted*.
- Marbac, M., Biernacki, C., & Vandewalle, V. 2013. *Model-based clustering for conditionally correlated categorical data*. Rapport de recherche RR-8232. INRIA.
- Maugis, C., Celeux, G., & Martin-Magniette, M-L. 2009. Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics & Data Analysis*, **53**(11), 3872–3882.

- McLachlan, G.J., & Krishnan, T. 1997. *The EM algorithm*. Wiley-Interscience, New York: Wiley Series in Probability and Statistics: Applied Probability and Statistics.
- McLachlan, G.J., & Peel, D. 2000. *Finite mixture models*. Wiley-Interscience, New York: Wiley Series in Probability and Statistics: Applied Probability and Statistics.
- Meila, M., & Jordan, M.I. 2001. Learning with mixtures of trees. *The Journal of Machine Learning Research*, **1**, 1–48.
- Muthén, B. 2008. Latent variable hybrids: Overview of old and new models. *Advances in latent variable mixture models*, **1**, 1–24.
- Qu, Y., Tan, M., & Kutner, M.H. 1996. Random Effects Models in Latent Class Analysis for Evaluating Accuracy of Diagnostic Tests. *Biometrics*, **52**(3), pp. 797–810.
- Reboussin, B.A., Song, E.Y., Shrestha, A., Lohman, K.K., & Wolfson, M. 2006. A latent class analysis of underage problem drinking: Evidence from a community sample of 16–20 year olds. *Drug and alcohol dependence*, **83**(3), 199–209.
- Reboussin, B.A., Ip, E.H., & Wolfson, M. 2008. Locally dependent latent class models with covariates: an application to under-age drinking in the USA. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **171**(4), 877–897.
- Richardson, S., & Green, P.J. 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **59**(4), 731–792.
- Robert, C.P. 2005. *Le choix bayésien: principes et pratique*. Springer France Editions.
- Robert, C.P., & Casella, G. 2004. *Monte Carlo statistical methods*. Springer Verlag.
- Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.
- Strauss, S.M., Rindskopf, D.M., Astone-Twerell, J.M., Des Jarlais, D.C., & Hagan, H. 2006. Using latent class analysis to identify patterns of hepatitis C service provision in drug-free treatment programs in the US. *Drug and alcohol dependence*, **83**(1), 15–24.

- Van Hattum, P., & Hoijtink, H. 2009. Market Segmentation Using Brand Strategy Research: Bayesian Inference with Respect to Mixtures of Log-Linear Models. *Journal of Classification*, **26**(3), 297–328.
- Vermunt, J.K. 2003. Multilevel latent class models. *Sociological methodology*, **33**(1), 213–239.
- Vermunt, J.K. 2007. Multilevel mixture item response theory models: an application in education testing. *Proceedings of the 56th session of the International Statistical Institute. Lisbon, Portugal*, 22–28.

A Dentistry clustering with the R package Clustericat

The R package `Clustericat` is available on Rforge website at the following url: https://r-forge.r-project.org/R/?group_id=1803. This section presents the code used to cluster the dentistry data set.

```
# Loading of the data set
> data("dentist")

# to define the parameters of the algorithm performing the estimation
# here 25 MCMC are performed with a stopping criterion equals to
# 200 successive iterations having not found a better model
> st <- strategycat(dentist, nb_init=25, stop_criterion=200)

# estimation of the model for a class number equal to 2.
# for the data set with five binary variables (modal)
> res <- clustercat(dentist, 2, modal=rep(2,5), st)

# presentation of the best model
> summary(res)

# presentation of the parameters of the conditional dependencies for the best model
> summary_dependencies(res)
```

```
# a plot summarizing the best model like Figure 6  
> plot(res)
```